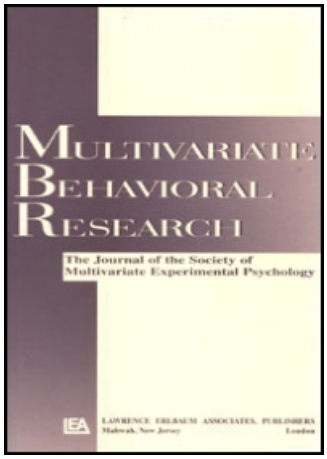
Influence of simple imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable

Bernaards, C.A.; Sijtsma, K.

Published in:
Multivariate Behavioral Research

Publication date:
2000

Document Version
Publisher's PDF, also known as Version of record



Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable

Coen A. Bernaards [a]; Klaas Sijtsma [b]
[a] Department of Methodology and Statistics FSW, Utrecht University.
[b] Department of Research Methodology FSW, Tilburg University.

Online Publication Date: 01 July 2000
To cite this Article: Bernaards, Coen A. and Sijtsma, Klaas (2000) 'Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable', Multivariate Behavioral Research, 35:3, 321 - 364
To link to this article: DOI: 10.1207/S15327906MBR3503_03
URL: http://dx.doi.org/10.1207/S15327906MBR3503_03

Multivariate Behavioral Research, 35 (3), 321-364

Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable

Coen A. Bernaards
Department of Methodology and Statistics FSW
Utrecht University

Klaas Sijtsma
Department of Research Methodology FSW
Tilburg University

This study deals with the influence of each of twelve imputation methods and two methods using the EM algorithm on the results of maximum likelihood factor analysis as compared with results obtained from the complete data factor analysis (no missing scores). Complete questionnaire rating scale data were simulated and, next, missing item scores were created under both ignorable and nonignorable nonresponse mechanisms. Next, imputation methods were used to fill the gaps and factor analysis was applied to both the original complete data and to the data sets including imputed scores. Each imputation method was implemented once with residual error and once without residual error. Also, one EM method estimated the factor loadings directly and the other estimated the complete data covariance matrix, which subsequently was factor analyzed. A design was analyzed with design factors Latent Trait Structure (technically called Mixing Configuration), Correlation Between Latent Traits, Nonresponse Mechanism, Percentage of Missingness, Sample Size, and Imputation Method. We found that, in general, methods that impute a score based on a respondent's mean score obtained from his/her observed item scores best recovered the factor loadings structure from the complete data. Moreover, for unidimensional data person mean methods with a residual error gave better results than the other imputation methods, either with or without a residual error component. For the EM methods a smaller design was analyzed. The conclusion was that both EM methods better recovered the complete data factor loadings than the imputation methods.

Introduction

Factor analysis is often used to study the structure of the item set in tests and questionnaires. A well known and difficult problem in data collection by means of tests and questionnaires is item nonresponse. Item nonresponse occurs if respondents are unable or reluctant to provide answers to one or more items or if they accidentally skip items, but at the same time produce answers to other items. In this article, we are concerned in particular with

item nonresponse for which the probability of not responding depends on the missing item score and cannot be explained by means of the completely observed variables (e.g. covariates). An example of such a phenomenon is when a respondent refuses to give an answer because a question is considered menacing to privacy (questions about one's sexual habits or income) or embarrassing (questions about the relationship with one's parents or children) and these opinions are not typical of the whole population or of particular subgroups. Nonresponse due to such a response mechanism is *nonignorable*.

Nonignorable item nonresponse may lead to a dataset which no longer is representative for the population of interest. For example, when people who earn high salaries more often do not answer a question about their income than people with lower salaries, the mean income based on the available data will be biased. Also, people who have a problematic relationship with their children may be inclined more than other people to skip questions in a questionnaire asking them about aspects of their interaction with their children; for example, frequency of reading to them before sleeping, helping them with their homework, and accompanying them to their sports events. These questions may induce nonresponse because they may be considered menacing. The statistical consequence again is biased estimates.

In other forms of nonresponse the missing responses may be a completely random phenomenon in the population at hand. In this first case, responses are missing completely at random (MCAR; Rubin, 1976). An example of MCAR is that respondents accidentally skip questions. Missingness may also occur as a random phenomenon within particular well-defined subgroups of the population but may be varying in degree between such subgroups. In this second case, responses are missing at random (MAR; Rubin, 1976). An example of MAR is that older respondents tend to accidentally skip more questions than younger respondents. Here, age is a covariate that explains differences between meaningful subgroups. The response mechanisms MCAR and MAR produce *ignorable* nonresponse.

The consequence of ignorable item nonresponse is that the sample of complete data cases is smaller than the original sample and, as a result, statistical estimates are less accurate but unbiased. Thus, one may argue that, all other things being equal (like sample size, percentage of missingness, etc.), ignorable nonresponse is less of a problem than nonignorable nonresponse because the latter problem in addition produces biased estimates. A practical problem in distinguishing between the two conditions is that the mechanism producing the nonresponse often is unknown (Huisman, 1998).

Only when the mechanism is known can the missingness be modeled adequately. In this study, we are dealing with data containing missing item

scores, in some datasets ignorable but in most datasets nonignorable, which was not modeled explicitly. Alternatively, for missing scores values were imputed according to simple methods and the resulting complete data were factor analyzed. Also, two versions of the EM algorithm were implemented. One was used for estimating factor loadings directly (henceforth, to be denoted EM-loadings) and the other for estimating a complete data covariance matrix (henceforth, to be denoted EM-covariances), after which the covariance matrix was factor analyzed. Because we knew the original complete data, it was possible to compare factor analysis results based on imputed data or based on the application of the EM methods with factor analysis results based on the complete data.

Item nonresponse does not include refusal of respondents to take part in the investigation, known as unit nonresponse, or attrition due to illness, moving to another city, and so on, known as experimental mortality. Thus, we consider the case when all respondents produced answers to at least some of the items, but not all respondents gave answers to all items.

Bernaards and Sijtsma (1999) discussed seven missing data methods in the context of factor analysis of rating scale data suffering from ignorable nonresponse (MCAR and MAR). Among these seven methods were five imputation methods, EM-loadings (Rubin & Thayer, 1982), and listwise deletion. It was found that the EM algorithm was superior to other missing data methods in the sense that the sum of squared differences between the factor loadings based on the complete data and the factor loadings based on the data with missing values imputed, was the smallest for EM. Listwise deletion was the worst method. Person mean (PM) was the best imputation method. Bernaards and Sijtsma (1999) noted that the existing literature (Cattell, 1978; Finkbeiner, 1979; Brown, 1983; Lee, 1986; Muthén, Kaplan & Hollis, 1987; also, see Liu & Rubin, 1998) on missing data methods in the context of factor analysis was much oriented towards dealing with missingness through maximum likelihood estimation rather than simple imputation. Moreover, methods based on regression analysis and principal components analysis were used. Huisman (1998) discussed several imputation methods in the context of scale construction and investigated the influence of these methods on the scale score, the reliability, and the scalability, but he did not deal with influence on factor analysis results.

In this article, we study the performance of twelve imputation methods for dealing mainly with nonignorable missing item scores in questionnaire data. For simulated questionnaire data containing missing item scores, the question was how well the use of these methods for producing complete data can lead to the reconstruction of the factors that resulted from the original complete data. This led to recommendations concerning the use of

imputation methods in practical questionnaire research where factor analysis of the data is envisaged but several item scores are missing. Also, results for ignorable item nonresponse were obtained. The EM methods (i.e., EM-loadings and EM-covariances) were used in a limited number of design cells so that results obtained by these relatively complex methods could be compared with results obtained by the simpler imputation methods.

The simulated data were the scores on a test or questionnaire consisting of ordered five-point rating scales (Likert items; Likert, 1932) and, for each simulee, scores on two covariates. Covariates were included because they are part of most practical studies and because covariate groups often systematically differ in their mean scores on the relevant variables. Thus, we used mean differences between covariate groups for generating our data. Knowledge of covariates was used in one imputation method although this was irrelevant for MCAR and insufficient for nonignorable missingness studied here. The rationale was that researchers often do not know which mechanism caused the missingness in their investigation and will use knowledge of covariates anyway. We used two binary covariates, which can be thought to represent, for example, gender and age group (say, young versus old).

Although many statistical models for analyzing questionnaire data assume that all items measure the same latent trait or underlying factor (unidimensionality), in practice responses frequently are the result of a combination of latent traits or underlying factors (multidimensionality). For example, responses to an item on introversion could partly be determined by language skills. Bernaards and Sijtsma (1999) found that data of higher dimensionality (more specifically, four-dimensional data) led to the same conclusions about the usefulness of missing data methods in factor analysis than two-dimensional data. Thus, in the present study only one- and two-dimensional data sets were simulated.

Complete simulated data matrices were generated by means of a multidimensional polytomous item response theory (IRT) model (Kelderman & Rijkes, 1994) and, next, subjected to factor analysis. Takane and De Leeuw (1987), Muraki and Carlson (1995), and McDonald (1997) discussed the relation between multidimensional IRT and factor analysis. Maximum likelihood factor analysis assumes normally distributed variables. Dolan (1994) and Muthén and Kaplan (1985, 1992) demonstrated that even with five-point rating scale data, factor analysis is not seriously affected by deviations from normality of the distributions of the variables.

Next, item scores were deleted both under MCAR and nonignorable missing data mechanisms (to be discussed later on in detail), and the resulting incomplete data matrix was then treated by subsequently applying one of the twelve imputation methods, which yielded twelve reconstructed data matrices. In a limited number of design cells, the two EM-based methods were applied

to the missing data problem. For each of the twelve imputed data matrices the same number of factors was extracted as for the complete data matrix. The number of factors equalled the number of latent traits used to generate the data. Bernaards and Sijtsma (1999) found that use of the exploratory eigenvalue-higher-than-1 criterion led to retention of the same number of factors, and concluded that using the prior knowledge of the number of factors was justified. Factor analysis results based on a complete data matrix could be compared with the results obtained under each of the imputation methods and the EM algorithms, and conclusions could be drawn with respect to the effectiveness of these methods for producing the correct results.

Method

Generating the Data

Item response theory was used for modeling the data generation process. We used IRT for defining the characteristics of k items from an imaginary questionnaire. We also chose a fixed number of latent traits (this is IRT terminology; using factor analysis terminology we would call them latent factors) and a multivariate distribution of these latent traits. Based on the item characteristics and the distribution of the latent traits, an IRT model (discussed in more detail later on) determined for each simulee his/her probabilities of responding in each of the categories of the rating scale of a particular item, and this was done for each item. This procedure generated a complete data matrix for N simulees and k items. Next, this data matrix was factor analyzed. The resulting *factor structure* was assumed to reflect the *latent trait structure*.

A doubly stochastic process (cf. Lord & Novick, 1968, pp. 29-30) was used for generating rating scale data. First, for N simulees two latent trait scores (or, equivalently, underlying factor scores) were randomly drawn from a bivariate normal distribution. Each simulee belonged to one of four covariate classes, which had different latent trait means. Second, for each simulee the scores on k rating scale items were drawn from a propensity distribution described by a multidimensional item response model (the number of dimensions equaled the number of latent traits; here this number was 2).

The generation of the latent traits used two binary covariates with scores for each person. In the population, the four possible combinations of covariate scores defined four equally large groups. When simulating data, for each of the N persons a combination of covariate scores was drawn with probability 1/4. Covariate class (0,0) had latent trait means (0,0); covariate classes (1,0) and (0,1) both had latent trait means (1,1); and covariate class (1,1) had latent

trait means (2,2). Covariates were only related to the latent trait means. The latent trait covariances were the same for each covariate group.

The multidimensional polytomous latent trait (MPLT) model (Kelderman & Rijkes, 1994) was used to generate the polytomous item scores. Items are indexed j (j = 1, ..., k) and each item has r + 1 ordered answer categories with scores x = 0, ..., r; here r = 4. The items measure a combination of latent traits according to some a priori known ratio per item. For example, ten items may measure latent trait A and latent trait B with weights 1 and 3, respectively, and the next ten items may measure these traits with weights 3 and 1, respectively. Latent traits are denoted by θ and scores by θ_{iq} with indices i (i = 1, ..., N) for identifying persons and indices q (q = 1, ..., s) for identifying traits; here, s = 2.

The scoring weights associated with the response categories are contained in the three-way array **B** with entries B_{jqx}. The scoring weights reflect the ratio by which item j measures the latent traits, and also can be interpreted as discrimination indices. Following Kelderman and Rijkes (1994), we maintain the terminology of scoring weights. The separation parameters for the categories associated with B_{jqx} are contained in the three-way array $\mathbf{\Psi}$ with indices Ψ_{jqx}. By choosing the scoring weights **B** appropriately, different models can be defined.

The MPLT model is of the form

$$(1) \qquad P(X_{ij}=x|\theta_{i1},...,\theta_{is})=\frac{\exp\left[\sum_{q=1}^{s}\left(\theta_{iq}-\Psi_{jqx}\right)B_{jqx}\right]}{\sum_{y=0}^{r}\left\{\exp\left[\sum_{q=1}^{s}\left(\theta_{iq}-\Psi_{jqy}\right)B_{jqy}\right]\right\}}.$$

The MPLT model requires that if B_{jqy} = 0, then Ψ_{jqy} = 0, to ensure uniqueness of the parameters.

Since the scoring weight array **B** in model 1 and the array of separation parameters $\mathbf{\Psi}$ were specified a priori, for each item the probabilities of response in each answer category could be calculated for each vector $\boldsymbol{\theta}$ drawn from the bivariate normal distribution. Next, for given vector $\boldsymbol{\theta}$ for each of the k items an outcome was drawn from a multinomial distribution with response probabilities (Equation 1) as calculated for the answer categories, resulting in k item responses for each of the N persons.

Defining and "Generating" Missings

Types of Item Nonresponse

Two types of nonignorable item nonresponse were defined. First, in a questionnaire with Likert (1932) scales scores may be missing especially in the higher answer categories. For example, an item testing preference towards the extreme political rightwing could have answer categories: no preference at all, no explicit dislike, weak preference, positive affection, and strong preference. Missings may occur in particular in the higher answer categories which represent socially the least desirable and the least politically correct answers.

Following this rationale, we generated missings such that higher answer categories had higher probabilities of nonresponse. Table 1 (upper panel, REF-HIGH) shows for each of the four covariate classes relative expected frequencies (REF's; 0 means never missing) which exhibit the trend that nonresponse is more likely for the higher answer categories. Because the literature did not reveal which REF's are the most realistic, we chose values that seemed reasonable.

Two aspects of the REF's are noteworthy. First, within a covariate class the trend of the REF's reflects that the probability of nonresponse increases toward the higher answer categories. Since the different covariate classes show different patterns of REF's of nonresponse, covariate classes explain some of the nonresponse but not all. Thus, within each covariate class, the probability of missingness is still related to the response category, and hence missingness and response category are not conditionally independent, as required by MAR. Second, for each covariate class the last column of Table 1 gives the expected percentage of missingness. To summarize, covariate classes differ in the percentage of missingness and in the pattern of missingness across the answer categories of a rating scale.

Moreover, as was discussed earlier covariance classes also define three different latent trait means, but these means do not completely coincide with the covariance classes [classes (1,0) and (0,1) have the same mean]. Thus, these different means also influence the data generation process, but the part of the nonresponse explained by the covariance classes is not exactly the same as the part explained by the latent trait means.

Second, missings also may arise especially in the extreme positive and the extreme negative answer categories. Again consider a five-point rating scale. The possible answers to the question "What is your annual income?" may be: low, below average, average, above average, and high. Both the categories "low" and "high" may tend to show missingness more often than

Table 1
Relative Expected Frequency (REF) of Nonresponse for Two Binary Covariates

		Response Categories					
Name	Covariates	0	1	2	3	4	% Missing
REF-HIGH	(0,0)	0	0	1	3	5	14
	(0,1)	0	1	3	5	5	22
	(1,0)	0	0	1	5	10	25
	(1,1)	0	1	5	10	10	40
REF-LOWHIGH	(0,0)	5	1	0	1	5	16
	(0,1)	5	3	0	3	5	21
	(1,0)	10	1	0	1	10	29
	(1,1)	10	3	0	3	10	34
REF-MCAR	(0,0)	1	1	1	1	1	25
	(0,1)	1	1	1	1	1	25
	(1,0)	1	1	1	1	1	25
	(1,1)	1	1	1	1	1	25

Note: The % Missing was obtained by adding, for each covariate class, the relative frequencies, and dividing this sum by the total sum across all four classes. For example, $14 = 100 \times (0 + 0 + 1 + 3 + 5)/65$ (first entry).

the three middle categories, because respondents with low or high incomes may be more secretive about this than other respondents.

We used this example to generate missingness with relatively high probability in the two extreme answer categories. The middle panel of Table 1 (REF-LOWHIGH) shows the REF's representing this type of missingness. The trends of the numbers in Table 1 reflect that the probability of nonresponse increases toward both extremes of the rating scale. The percentage of missingness varies across covariate classes. Also, covariate classes explain some but not all of the missingness and the amount explained does not exactly coincide with that explained by the latent trait mean. As in the previous case missingness is nonignorable.

Finally, as a benchmark for the other two cases the lower panel of Table 1 (REF-MCAR) shows REF's typical of MCAR: The REF's have no relation to covariate class (if they would, we would have MAR) and, moreover, no relation to answer category (if they would, missingness would be nonignorable).

Generating Missing Item Scores

Next, we illustrate the procedure of creating missings and keeping the percentage of missingness in a sample at a preset level of, say, 20 percent. The artificial data matrix in panel A of Table 2 shows for 15 simulees scores ranging from 0 to 4 on six five-point rating scales, generated using the MPLT model and a θ with normal distribution having mean 0 and variance 2½. Assume that these data stem from covariate class (0,0) (mean θs of 0) and that missingness occurs on each of the items according to relative frequency 00135 (upper panel of Table 1). We used SPlus (Becker, Chambers, & Wilks, 1988) for simulating this missingness pattern for the data in panel A of Table 2 which resulted in the data matrix in panel D.

Panel B in Table 2 shows the matrix which has for each score from panel A the corresponding relative frequency (thus, panel B has entries 0, 1, 3, and 5). Each of these entries is divided by the sum of the 90 entries in panel B (this sum equals 128) and this yields the probabilities in panel C. A missing is created by sampling (without replacement) 18 times (20 percent out of 90 scores) from a multinomial distribution with 90 categories and the probabilities in panel C. This way, exactly 20 percent missings are created; the end result is shown in panel D.

Imputation Methods and EM Methods

Imputation Methods

Imputation methods estimate the missing score and then impute this estimate. The advantage of imputation over other missing data procedures is that standard complete-data methods can be used for further data analysis.

The following imputation methods were implemented.

1. *Overall Mean Imputation (OM)* replaces the missing values by the mean across all observed item responses in the data matrix.

2. *Person Mean Imputation (PM)* for each person separately calculates the mean across all of his/her available item reponses, and imputes this mean for each missing value for that particular person.

Table 2
Example of Generating Missings

Panel																							
A: Complete Data						B: REF						C: Probabilities						D: Incomplete Data					
2	1	1	2	1	1	1	0	0	1	0	0	1/128	0	0	1/128	0	0	2	1	1	2	1	1
2	2	2	1	1	1	1	1	1	0	0	0	1/128	1/128	1/128	0	0	0	2	2		1	1	1
3	4	4	3	4	4	3	5	5	3	5	5	3/128	5/128	5/128	3/128	5/128	5/128	3	4	4	3		
3	2	1	2	3	2	3	1	0	1	3	1	3/128	1/128	0	1/128	3/128	1/128	3	2	1	2	3	2
0	1	2	0	1	2	0	0	1	0	0	1	0	0	1/128	0	0	1/128	0	1	2	0	1	2
3	3	2	2	3	3	3	3	1	1	3	3	3/128	3/128	1/128	1/128	3/128	3/128	3		2	2		
3	2	3	2	2	2	3	1	3	1	1	1	3/128	1/128	3/128	1/128	1/128	1/128		2		2	2	2
3	2	2	1	2	3	3	1	1	0	1	3	3/128	1/128	1/128	0	1/128	3/128	3	2	2	1	2	
2	3	2	3	3	1	1	3	1	3	3	0	1/128	3/128	1/128	3/128	3/128	0		3	2			1
2	1	2	2	2	2	1	0	1	1	1	1	1/128	0	1/128	1/128	1/128	1/128	2	1		2	2	2
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2	2	3	3	2	2	1	1	3	3	1	1	1/128	1/128	3/128	3/128	1/128	1/128	2		3	3	2	2
2	1	2	2	2	2	1	0	1	1	1	1	1/128	0	1/128	1/128	1/128	1/128	2	1	2	2	2	
3	3	4	4	3	2	3	3	5	5	3	1	3/128	3/128	5/128	5/128	3/128	1/128	3	3	4		3	
1	1	2	1	2	1	0	0	1	0	1	0	0	0	1/128	0	1/128	0	1	1		1	2	1

3. *Conditional Mean Imputation (CM)* calculates the mean across all available item scores of all persons within the same covariate class, and imputes this value for all missing scores within the same covariate class.

4. *Item Mean Imputation (IM)* calculates for each item the mean across all available scores, and imputes this mean for that particular item.

5. *Two-Way Imputation (TW)* calculates across available scores the overall mean, the mean for item *j* and the mean for person *i*, and imputes IM + PM - OM for missing observation (*i*, *j*).

6. *Corrected Item Mean Substitution (CIMS)* (Huisman, 1998) imputes for an unobserved itemscore in row *i* and column *j* the following value, x_{ij}, based on available scores,

$$(2) \qquad x_{ij} = \left(\frac{\sum_{j \in obs(i)} x_{ij}}{\sum_{j \in obs(i)} \bar{x}_{.j}} \right) \bar{x}_{.j} = \left(\frac{\mathrm{PM}_i}{\frac{1}{\#obs(i)} \sum_{j \in obs(i)} \mathrm{IM}_j} \right) \mathrm{IM}_j ,$$

where $\bar{x}_{.j}$ is the mean across all observed scores on item *j*, *obs*(*i*) is the set of all observed scores for respondent *i*, and #*obs*(*i*) is the number of observed scores for respondent *i*; PM_i is the person mean for respondent *i* and IM_j is the item mean for item *j*. The ratio between brackets thus estimates a multiplication factor for respondent *i* as PM_i divided by the mean of the item means across all observed items for respondent *i*. This scalar is higher than 1 for respondents scoring higher than average and lower than 1 for respondents scoring lower than average. Each missing item score of respondent *i* is then replaced by the mean of item *j* times the multiplication factor. CIMS thus takes the 'ability' of the respondent into account by imputing a higher score the higher the scores on the completed items.

Based on the incomplete data matrix in panel D of Table 2, Table 3 contains for each of the six imputation methods the imputed data matrix. It may be noted that for only one group (one covariate class) methods OM and CM produce the same imputed scores.

For all six imputation methods a second version was implemented which adds a random draw from a normal distribution with mean zero and residual variance. The reason for adding a residual normal deviate to the imputed values was to include sampling fluctuation. Thus, results may be more realistic compared with results based on imputing an error-free mean. For example, for method OM the difference is calculated between the data matrix and a matrix which consists completely of imputed values (all elements equal to OM). Next, the sample residual variance is calculated

Table 3

Imputed Data Matrices for Six Imputation Methods (OM, CM, PM, IM, TW, CIMS) Based on Panel D from Table 2

OM = CM						PM							IM						TW							CIMS							
												PM												PM						PM	DEN_i[a]
2	1	1	2	1	1	2	1	1	2	1	1	1.3	2	1	1	2	1	1	2	1	1	2	1	1	1.3	2	1	1	2	1	1	1.3	1.8
2	2	**1.8**	1	1	1	2	2	**1.4**	1	1	1	1.4	2	2	**2.1**	1	1	1	2	2	**1.7**	1	1	1	1.4	2	2	**1.7**	1	1	1	1.4	1.7
3	4	4	3	**1.8**	**1.8**	3	4	4	3	**3.5**	**3.5**	3.5	3	4	4	3	**1.8**	**1.4**	3	4	4	3	**3.5**	**3.1**	3.5	3	4	4	3	**3.4**	**2.6**	3.5	1.9
3	2	1	2	3	2	3	2	1	2	3	2	2.2	3	2	1	2	3	2	3	2	1	2	3	2	2.2	3	2	1	2	3	2	2.2	1.8
0	1	2	0	1	2	0	1	2	0	1	2	1.0	0	1	2	0	1	2	0	1	2	0	1	2	1.0	0	1	2	0	1	2	1.0	1.8
3	**1.8**	2	2	**1.8**	**1.8**	3	**2.3**	2	2	**2.3**	**2.3**	2.3	3	**1.8**	2	2	**1.8**	**1.4**	3	**2.3**	2	2	**2.4**	**1.9**	2.3	3	**2.2**	2	2	**2.2**	**1.7**	2.3	1.9
1.8	2	**1.8**	2	2	2	**2.0**	2	**2.0**	2	2	2	2.0	**2.0**	2	**2.1**	2	2	2	**2.2**	2	**2.3**	2	2	2	2.0	**2.4**	2	**2.5**	2	2	2	2.0	1.7
3	2	2	1	2	**1.8**	3	2	2	1	2	**2.0**	2.0	3	2	2	1	2	**1.4**	3	2	2	1	2	**1.6**	2.0	3	2	2	1	2	**1.5**	2.0	1.9
1.8	3	2	**1.8**	**1.8**	1	**2.0**	3	2	**2.0**	**2.0**	1	2.0	**2.0**	3	2	**1.6**	**1.8**	1	**2.2**	3	2	**1.8**	**2.0**	1	2.0	**2.3**	3	2	**1.8**	**2.1**	1	2.0	1.8
2	1	**1.8**	2	2	2	2	1	**1.8**	2	2	2	1.8	2	1	**2.1**	2	2	2	2	1	**2.1**	2	2	2	1.8	2	1	**2.2**	2	2	2	1.8	1.7
0	0	0	0	1	0	0	0	0	0	1	0	0.2	0	0	0	0	1	0	0	0	0	0	1	0	0.2	0	0	0	0	1	0	0.2	1.8
2	**1.8**	3	3	2	2	2	**2.4**	3	3	2	2	2.4	2	**1.8**	3	3	2	2	2	**2.4**	3	3	2	2	2.4	2	**2.4**	3	3	2	2	2.4	1.8
2	1	2	2	2	**1.8**	2	1	2	2	2	**1.8**	1.8	2	1	2	2	2	**1.4**	2	1	2	2	2	**1.4**	1.8	2	1	2	2	2	**1.4**	1.8	1.9
3	3	4	**1.8**	3	**1.8**	3	3	4	**3.3**	3	**3.3**	3.3	3	3	4	**1.6**	3	**1.4**	3	3	4	**3.1**	3	**2.9**	3.3	3	3	4	**2.7**	3	**2.4**	3.3	1.9
1	1	**1.8**	1	2	1	1	1	**1.2**	1	2	1	1.2	1	1	**2.1**	1	2	1	1	1	**1.5**	1	2	1	1.2	1	1	**1.5**	1	2	1	1.2	1.7

Overall Mean	Item Means	Item Means	OM	Item Means
1.8	2.0 1.8 2.1 1.6 1.8 1.4	2.0 1.8 2.1 1.6 1.8 1.4	1.8	2.0 1.8 2.1 1.6 1.8 1.4

[a] DEN_i is the denominator of Equation 2 for person *i*.

across all available differences (cells containing missing item scores are omitted). Finally, each missing item response in the observed data matrix is substituted by the sum of OM and a draw from a normal distribution with mean zero and residual variance. The imputation methods with residual variance are denoted by OM-E, PM-E, CM-E, IM-E, TW-E, and CIMS-E. Thus, in total 12 imputation methods were implemented.

An advantage of most imputation methods over other missing data methods, for example, the EM algorithm, is their simplicity. OM is probably the simplest method, yet the most naieve because the mean is taken over all latent traits and all classes of covariates. CM partly alleviates this drawback by taking the mean within classes of covariates. However, as with OM multidimensionality of the data is ignored. Method IM corrects for multidimensionality but not for classes of covariates. Thus, it is difficult to predict whether IM performs better than CM.

PM takes the mean over the smallest meaningful group of item responses. Thus, PM may lead to the most meaningful imputed score of all methods, because each respondent is treated uniquely and an imputed value is based on correlated items. Hence we expect (see also Bernaards and Sijtsma, 1999) that in unidimensional data PM will give better results than IM, CM, and OM. In two-dimensional data, the advantage of using correlated items is lost to some degree, because the item weights of 1 used in calculating the person mean clearly are less appropriate than when unidimensionality applies. It may be noted that items weights of 1 would even be more inappropriate when some of the inter-itemcorrelations were negative (this does not happen in the present research). Thus, for two-dimensional data the relative performance of PM is more difficult to predict. Moreover, in all cases the imputed PM score may be subject to higher uncertainty than imputed scores based on other methods because PM uses the smallest subset of available scores.

TW is based on the two-way layout used in ANOVA models by imputing a row-effect plus a column-effect minus the overall effect. This method was suggested to us by D.B. Rubin (personal communication, November 21, 1997). TW corrects for multidimensionality via IM, for ability per person via PM, and for the overall effect via OM. Hence, TW may be expected to better recover the matrix of factor loadings based on the complete data than IM, PM, and OM separately. CIMS, like TW, corrects for multidimensionality via IM and for ability per person via PM.

Usually, calculation of a mean will not generate an integer. The imputed mean values thus are not “valid” scores as would have been observed had these scores not been missing. Reals rather than integers were used, however, because the present study was concerned only with results from

factor analysis and not with the imputed scores themselves. Moreover, rounding of imputed scores to the nearest integer would introduce undesirable additional error into the data.

EM Methods

Bernaards and Sijtsma (1999) found that the EM algorithm (version EM-loadings) recovered the complete data factor loadings considerably better than simple imputation methods. Thus, it was expected that EM-loadings would also perform better in the present study. Although the relatively complex algorithm may not be easily accessible to many practical researchers and, moreover, this study concentrated on imputation methods, EM-loadings was implemented in a limited part of the design. This way, it could be checked whether EM-loadings also performed better than other methods when missingness is nonignorable. In addition, EM-covariances was implemented both as a competitor to EM-loadings and also for comparison with the simpler imputation methods. Since Bernaards and Sijtsma (1999) found that listwise deletion by far gave the worst results, this method was left out of the present study.

EM-loadings handles the factor scores from factor analysis as missing. Initially, random values are substituted for the missing data. In the E step, the missing values are updated given the factor scores, and the expected value of the covariance matrix given the factor loadings is calculated. In the M step the factor loadings and factor scores are updated based on the current estimate of the covariance matrix. These two steps are re-iterated until convergence of the likelihood occurs. The EM implementation used here is described in detail in the Appendix of Bernaards and Sijtsma (1999).

Assuming that the sample originates from a multivariate normal distribution, EM-covariances estimates the population covariance matrix based on the data including missing item scores (Little & Rubin, 1987; Schafer, 1997). Initially, parameters are estimated using listwise deletion. In the E step, expectations and (co)variances are calculated for the missing data given the observed values and the current parameter estimates. In the M step, the parameter estimates are updated based on the current expectations and (co)variances of the data including missing item scores. Iteration between the E step and the M step continues until convergence of the parameters. This implementation is described in detail in the Appendix.

Theoretically, EM-loadings is expected to better estimate the loadings than EM-covariances because EM-loadings directly estimates the loadings and EM-covariances uses one additional estimation step (which is the estimation of the complete data covariance matrix) prior to factor analysis.

This estimation step may introduce additional error in the estimates of the loadings.

Summary

To summarize, this study differed from the study by Bernaards and Sijtsma (1999) because: (a) listwise deletion was left out; (b) two versions of EM were studied to check whether they also were the best methods when missingness was nonignorable; (c) methods TW and CIMS were added as new and promising imputation methods; and (d) for each imputation method a version with a residual error was considered.

Other methods, such as multiple imputation (Rubin, 1987) and computer-intensive methods as described by Tanner (1996), are more difficult to implement and to understand for practical researchers and were not considered here. Acock (1997) gives an elementary introduction to missing data methods used with social science data. Also see Little and Rubin (1987) for an elaborate treatment of missing data methods. Other sources on missing data methods are, for example, Little and Rubin (1989), Rubin (1991), Little and Schenker (1995), Rubin (1996), and Schafer (1997).

Performance of Methods

Bernaards and Sijtsma (1999) used the statistics Tucker's ϕ and D^2 for evaluating the discrepancy between a factor loadings matrix based on the complete data and a factor loadings matrix based on the data to which a missing data method had been applied. Tucker's ϕ measures similarity of direction within pairs of vectors. For example, ϕ is calculated between the first vector of loadings based on the complete data and the first vector of loadings based on the data including imputed values; between the second vector of loadings based on the complete data and the second vector of loadings based on the data with imputed scores; and so on. ϕ lies between 0 and 1, where 1 means that two vectors point in exactly the same direction. For $\phi > 0.85$ two vectors of loadings have the same interpretation (Ten Berge, 1977; Niesing, 1997). Bernaards and Sijtsma (1999) found that in their study all ϕs were higher than 0.98, and concluded that the direction of the loadings was not influenced by the application of the missing data methods. Thus, we will not consider ϕ in the present article.

The second statistic used by Bernaards and Sijtsma (1999) for evaluating discrepancy between matrices of loadings, D^2, was the sum of squared differences between the two matrices of factor loadings, both based on all retained factors, and corrected for the number of retained factors. The

smaller D^2, the more similar the two loadings matrices. From loadings matrices **X** and **Y**, D^2 can be calculated through

$$D^2 = tr\left[(\mathbf{X}-\mathbf{Y})^T(\mathbf{X}-\mathbf{Y})\right]/m = tr\left[(\mathbf{X}-\mathbf{Y})(\mathbf{X}-\mathbf{Y})^T\right]/m,$$

where m is the number of factors retained. D^2 will be used in the present article, with **X** representing the complete data loadings matrix and **Y** representing the loadings matrix based on the data including imputed scores. For EM-loadings, **Y** was the direct EM estimate of the loadings and for EM-covariances **Y** was the loadings matrix based on the EM estimated covariance matrix.

The second method used to monitor performance is based on testing the equality of two covariance matrices. Let $\mathbf{\Sigma}_c$ denote the k-dimensional, unknown, true covariance matrix of the complete data. Analogously, $\mathbf{\Sigma}_i$ (subscript i for "imputed"; not for indexing persons) denotes the k-dimensional true covariance matrix of the imputed data resulting from a particular missing data method. Since we are concerned with the recovery of complete data factor loadings from data to which a missing data method has been applied, we first calculate the estimated covariance matrices $\hat{\mathbf{\Sigma}}_c$ and $\hat{\mathbf{\Sigma}}_i$ as follows. Let $\hat{\mathbf{\Sigma}}_c = \hat{\mathbf{\Lambda}}_c \hat{\mathbf{\Lambda}}_c' + \hat{\mathbf{\Psi}}_c$, where $\hat{\mathbf{\Lambda}}_c$ is the matrix of estimated factor loadings based on the complete data and where $\hat{\mathbf{\Psi}}_c$ contains the corresponding uniquenesses. Analogously, $\hat{\mathbf{\Sigma}}_i = \hat{\mathbf{\Lambda}}_i \hat{\mathbf{\Lambda}}_i' + \hat{\mathbf{\Psi}}_i$ denotes the estimated covariance matrix based on the data with imputed scores using a particular missing data method.

Test statistics for the null hypothesis $\mathbf{\Sigma}_c = \mathbf{\Sigma}_i$ are based on the eigenvalues $\gamma_1, \ldots, \gamma_k$ of the equation (Anderson, 1984, pp. 422-424)

$$\begin{aligned} &\hat{\mathbf{\Sigma}}_i \mathbf{x} = \hat{\mathbf{\Sigma}}_c \gamma \mathbf{x} \\ &\Leftrightarrow \hat{\mathbf{\Sigma}}_c^{-1} \hat{\mathbf{\Sigma}}_i \mathbf{x} = \gamma \mathbf{x}, \end{aligned} \tag{3}$$

where **x** is the eigenvector of $\hat{\mathbf{\Sigma}}_c^{-1}\hat{\mathbf{\Sigma}}_i$ corresponding to eigenvalue γ. When factor analysis of the data with imputed scores perfectly recovers the loadings from complete data factor analysis, the product $\hat{\mathbf{\Sigma}}_c^{-1}\hat{\mathbf{\Sigma}}_i$ results in all eigenvalues equal to 1. In practice, however, the eigenvalues tend to be scattered around 1. The magnitude of this spread indicates how much the two matrices differ from one another. To determine this spread, several test statistics have been proposed in the literature, see Anderson (1984) and Stevens (1992) for overviews. We prefered to use “simple” descriptive functions over complicated statistical tests in order to determine the spread of the eigenvalues. Examples of such descriptive functions are the mean of

the eigenvalues, the product of the eigenvalues, $\Pi\gamma$, which equals the matrix determinant, and the ratio $\max\gamma/\min\gamma$ of the eigenvalues.

How should these functions be interpreted? The product of the eigenvalues based on data obtained from k items equals the volume of the k-dimensional hyperparallelepiped spanned by the columns of the matrix $\hat{\boldsymbol{\Sigma}}_c^{-1}\hat{\boldsymbol{\Sigma}}_i$. The mean of the eigenvalues is the squared length of the interior diagonal of the hyperparallelepiped. The ratio $\max\gamma/\min\gamma$ does not have a clear geometrical interpretation.

Unfortunately, it is unclear which measure of similarity should be used. For example, consider the similarity of 3×3 correlation matrices $\mathbf{R}_i$ and $\mathbf{R}_c$. Because a correlation matrix has a trace equal to its dimension (here 3), the mean of the eigenvalues is equal to 1. This renders the mean useless. Next, assume that Imputation Method 1 results in eigenvalues 1.5, 0.75, and 0.75, and that Imputation Method 2 results in eigenvalues 1.4, 0.95, and 0.65. Then the volume of the parallelepiped (product of eigenvalues) based on Imputation Method 1 equals 0.84 and the volume based on Imputation Method 2 equals 0.86. However, $\max\gamma/\min\gamma$ based on Imputation Method 1 is 2 and $\max\gamma/\min\gamma$ based on Imputation Method 2 is 2.15. When the two correlation matrices are equal, $\Pi\gamma = \max\gamma/\min\gamma = 1$. Thus, it is difficult to decide which method to use because the results from the two measures of similarity can be contradictory. If both methods would always yield the same ordering of imputation methods, then one might decide to use only one.

The eigenvalues approach has two advantages over the squared distance between loadings D^2. The eigenvalues approach (a) is invariant under affine linear transformations, that is, transformations of the type $\mathbf{A}\hat{\boldsymbol{\Sigma}}_c + \mathbf{B}$ where $\mathbf{A}$ is a $k \times k$ non-zero matrix and $\mathbf{B}$ is any $k \times k$ matrix; and (b) takes the uniquenesses (errors of the variances) into account.

In addition to D^2, in this study the measures $\max\gamma/\min\gamma$ and $\Pi\gamma$ were used. When an imputation method performs well, the ratio $\max\gamma/\min\gamma$ and the product $\Pi\gamma$ both are close to 1. The ratio $\max\gamma/\min\gamma$ and the product $\Pi\gamma$ together indicate the deformation of the unithypercube which would be the result of $\hat{\mathbf{R}}_c^{-1}\hat{\mathbf{R}}_i$ in the case of perfect reconstruction of the complete data correlation matrix by the matrix based on data including imputed scores.

Design of Simulation Study

Choices relevant to the simulation design are summarized in Table 4. The upper panel has the design factors with varying levels and the lower panel has the design characteristics which were fixed throughout the investigation. This section discusses the design in more detail.

Latent Traits

The two latent traits θ_1 and θ_2 (Table 4) were assumed to be bivariate normally distributed (Table 4) with mean varying across covariate classes, as described earlier. Following Bernaards and Sijtsma (1999), correlations between traits were 0, 0.24, and 0.5, respectively (Table 4). Equation 4 provides the covariance matrices corresponding with the correlations between the traits.

$$(4) \qquad \begin{pmatrix} 2.5 & 0 \\ 0 & 2.5 \end{pmatrix} \begin{pmatrix} 2.5 & 0.6 \\ 0.6 & 2.5 \end{pmatrix} \begin{pmatrix} 2.5 & 1.25 \\ 1.25 & 2.5 \end{pmatrix}$$

Item Characteristics

Across the design, the number of items was fixed at $k = 20$ (Table 4), divided into two groups of ten items each. All items had five answer categories (Likert scales; Likert, 1932), scored $x = 0, ..., 4$ (Table 4). Each group of ten items consisted of three items with high score mode, four items with medium score mode, and three items with low score mode. Specifically, the separation parameters (MPLT model; see Equation 1) of the items were fixed (Table 4) such that for the first three items in a group of ten the mode was 3; for the next four items the mode was 2; and for the last three items the mode was 1.

In the two-dimensional datasets, each item measured both traits in a given ratio (or mixture; henceforth, abbreviated Mix; see Table 4). In configuration Mix 3:1, the first ten items measured the latent traits in the ratio 3:1, and the last ten items measured the latent traits in the reversed ratio 1:3. The interpretation of these ratio's is the following. The scoring weights B_{jqx} of the MPLT model (Equation 1) depend on item score x. Specifically, for higher x a higher weight B_{jqx} indicates a stronger dependence on latent trait θ_q (Kelderman & Rijkes, 1994). Thus, if an item measures both θ_1 and θ_2 with ratio 1:3, for a particular x this is reflected by the ratio of the weights B_{jqx} for θ_1 and θ_2. For example, in Table 5 (first panel, third and fourth row), for each item j ($j = 1, ..., 10$) for score 0, $B_{j10} = 1$ and $B_{j20} = 3$; for score 1, $B_{j11} = 2$ and $B_{j21} = 6$; and so on.

In Mix 1:0, the first ten items exclusively measured the first latent trait (Table 5, second panel; for θ_2 all Bs are zero), and the last ten items exclusively measured the second latent trait (Table 5, second panel; for θ_1 all Bs are zero). Unidimensionality was represented by Mix 1:1, in which all items measured both latent traits with the same pairwise weights (Table 5, third panel).

Table 4
Design Factors and Design Characteristics Relevant to the Simulation Study (First Panel Contains Design Factors with Varying Levels. Second Panel Contains Design Characteristics Fixed Throughout the Design.)

Design Factor	Levels
Correlation between latent traits	0, 0.24, 0.5
Scoring weights **B**	Mix 3:1, Mix 1:1, Mix 1:0
Percentage missingness	5, 10, 20
Missing Data Methods	Overall Mean (+error) Conditional Mean (+error) Item Mean (+error) Person Mean (+error) Two-Way imputation (+error) CIMS (+error) EM algorithm (two versions)
Sample size	100, 500
Relative expected frequency of nonresponse	REF-HIGH, REF-LOWHIGH, MCAR (see Table 1)
Performance of Method	D^2, $\Pi\gamma$, $\max\gamma/\min\gamma$
Design Characteristics	**Fixed**
Number of latent traits	2; bivariate normal, variance 2.5
Number of items	20
Number of answer categories	5 (ordered scores 0, ..., 4)
Separation parameters $\boldsymbol{\Psi}$	fixed per item
Extraction method	Maximum likelihood
Method of rotation	Varimax

Simulation of Data Matrices

The final step in generating item scores was determining for each combination of a simulee (defined by the θ-values) and each item (defined by its parameters $\boldsymbol{\Psi}$ and weights **B**) the probabilities of responding in each of the five answer categories. Comparison of these probabilities with draws from a uniform distribution led to an actual item score. For all $N \times k$ combinations together this yielded a complete data matrix.

Table 5
Scoring Weights **B** for MPLT Model

Mix	Latent Trait	Item Numbers	**B**
3:1	θ_1	1, ..., 10	3, 6, 9, 12, 15
	θ_2	1, ..., 10	1, 2, 3, 4, 5
	θ_1	11, ..., 20	1, 2, 3, 4, 5
	θ_2	11, ..., 20	3, 6, 9, 12, 15
1:0	θ_1	1, ..., 10	1, 2, 3, 4, 5
	θ_2	1, ..., 10	0, 0, 0, 0, 0
	θ_1	11, ..., 20	0, 0, 0, 0, 0
	θ_2	11, ..., 20	1, 2, 3, 4, 5
1:1	θ_1	1, ..., 20	1, 2, 3, 4, 5
	θ_2	1, ..., 20	1, 2, 3, 4, 5

Generating Missings

The procedure of creating missings in the complete data matrices was described in the section entitled "Defining and 'Generating' Missings" and Table 1. Following Bernaards and Sijtsma (1999), the percentages of missingness in the datamatrices were 5, 10, and 20, respectively (Table 4).

Imputation Methods

Twelve imputation methods were implemented; see the section entitled "Imputation Methods and EM Methods" and Table 4. Each method was used separately for imputing scores in the empty spaces of each data matrix. For each data matrix this yielded twelve different versions with imputed scores.

EM-Algorithms

Using maximum likelihood factor analysis, the EM-algorithm was used for handling the missing scores (Table 4). EM-loadings was described in detail by Bernaards and Sijtsma (1999). EM-covariances is described in detail in the Appendix.

Maximum Likelihood Factor Analysis

Bernaards and Sijtsma (1999) found that D^2 between complete and incomplete loadings based on principal components factor analysis was almost indistinguishable from D^2 between complete and incomplete loadings based on maximum likelihood factor analysis. Because from a mathematical point of view, maximum likelihood extraction is to be preferred over principal components factor analysis, the former method was used here (Table 4).

Sample Size

The sample sizes were fixed at 100 and 500 (Table 4). This seems to be well in agreement with studies of factor analysis performed on simulated data. Dolan (1994) used 200, 300, and 400 simulated respondents; Muthén and Kaplan (1985) were interested in large sample properties and used a sample size of 1000. Muthén and Kaplan (1992) extended their 1985 study by using sample sizes of 500 and 1000.

Varimax Rotation

All two-dimensional factor solutions were subjected to varimax rotation (Table 4). Varimax rotation is the most popular rotation method among practical researchers (it also is the default in SPSS, 1989); factor solutions are rotated to simple structure to facilitate interpretation; see, for example, Stevens (1992, p.380). One might argue that when latent traits are correlated, oblique rotation is more appropriate. We argued, however, that practical researchers often do not know whether factors are correlated and generally will use orthogonal rotation to simple structure.

It is not obvious whether the factor loadings based on a data matrix including imputed scores have the same alignment as factor loadings based on the complete data. Bernaards and Sijtsma (1999) performed orthogonal procrustes rotation of the loadings matrix based on the data including imputed scores towards the complete data factor loadings. They concluded that D^2 was not affected by this additional rotation, and hence the factor loadings based on the data including imputed scores could not be distinguished from the complete data target. Thus, the problem of alignment was not investigated any further here.

The number of factors retained depended on the number of latent traits used to generate the data. For the two-dimensional cases, Mix 3:1 and Mix 1:0, two factors were extracted. For the unidimensional case Mix 1:1, one factor was extracted.

Nonresponse Mechanism

As described earlier, three cases of missingness were studied (Table 4): missingness in higher categories of a rating scale (REF-HIGH; see Table 1); missingness in both low and high answer categories (REF-LOWHIGH; see Table 1); and missing completely at random (REF-MCAR; see Table 1).

Further Decisions

Performance of the imputation methods and the EM algorithms was evaluated using statistics D^2, $\Pi\gamma$, and $\max\gamma/\min\gamma$ (Table 4). Finally, 50 replications were carried out in each cell.

Design

For imputation methods, the design thus had 3 (mixing configurations) × 2 (sample size) × 3 (percentage missingness) × 3 (nonresponse mechanism) × 3 (correlation between traits) = 162 cells. For EM methods, due to long computation time and because Bernaards and Sijtsma (1999) found that, compared with imputation methods, EM-loadings always performed better, only part of the design was analyzed here. We chose a limited design consisting of the design factors mixing configuration (Mix 3:1, Mix 1:0, and Mix 1:1), percentage missingness (5, 20), nonresponse mechanism (REF-HIGH and REF-LOWHIGH), and correlation between traits (0, 0.24, 0.5). This yielded a design with 36 cells. Sample size was fixed at 100.

Results

Imputation Methods

We start this section with some preliminary decisions that serve to simplify the discussion of the results. Next, the results are discussed for each of the relevant factors from the design. For each design factor, first results for two-dimensional datasets are discussed, followed by results for unidimensional datasets.

In general, the ratio $\max\gamma/\min\gamma$ only changed substantially with the percentage of missingness and the sample size. For each imputation method separately, $\max\gamma/\min\gamma$ was approximately constant across all nonresponse mechanisms, correlations between traits, and mixing configurations. Thus, the ratio $\max\gamma/\min\gamma$ was not informative about possible discrepancies between loadings matrices for different imputation methods and,

consequently, may miss important results. Therefore, we confined the discussion of the results to D^2 and $\Pi\gamma$.

In almost all cells, the imputation methods IM-E, CM-E, and OM-E had the highest $\overline{\Pi\gamma}$ and $\overline{D^2}$ (mean) values. Because of this clearcut result, these methods were left out of further discussion of the results.

We now discuss the results for the imputation methods TW, CIMS, PM, IM, CM, OM, TW-E, CIMS-E, and PM-E with respect to correlation between traits, nonresponse mechanism (see Table 1), mixing configuration of the latent traits (see Table 5), and sample size, respectively. The Tables 6 through 8 give results for $\overline{\Pi\gamma}$ and $\overline{D^2}$. To save space, standard deviations $s(\Pi\gamma)$ and $s(D^2)$were not tabulated but the results are discussed along with the results for $\overline{\Pi\gamma}$ and $\overline{D^2}$. Entries ">>" in the columns labeled $\overline{\Pi\gamma}$ were higher than 20, and entries "<<" were lower than 1/20. The exact numerical values were too extreme to be of much importance and thus were left out. The imputation methods IM, CM, and OM had the highest number of extreme values for $\overline{\Pi\gamma}$.

Correlation Between Traits

For two-dimensional datasets (Mix 1:0 and Mix 3:1), for all imputation methods higher correlation between traits led to closer correspondence ($\overline{\Pi\gamma}$ closer to 1) between loadings matrices for complete data and imputed data, respectively. Also, all imputation methods showed smaller spread [$s(\Pi\gamma)$] as the correlation increased (not tabulated). This is illustrated in Figure 1, which contains for imputation method TW histograms of all 20 eigenvalues over 50 replications, for correlation 0 (upper panel), correlation 0.24 (middle panel) and correlation 0.5 (lower panel); and keeping other design factors fixed at Mix 3:1; sample size 100, REF-LOWHIGH, and 20 percent missings.

For two-dimensional datasets (Mix 1:0 and Mix 3:1), methods TW, CIMS, PM, TW-E, CIMS-E, and PM-E always showed a decrease in $\overline{D^2}$ and $s(D^2)$ [i.e., closer correspondence between loadings matrices; $s(D^2)$ not tabulated] when the correlation increased. For the other methods (IM, CM, OM), $\overline{D^2}$ and $s(D^2)$ did not change as the correlation increased. The six imputation methods that showed a decrease in $\overline{D^2}$ and $s(D^2)$ all used the person mean whereas the other three methods did not. Since latent traits were more similar when the correlation increased, the estimates of imputed scores based on methods using the person mean tended to be less biased and, therefore, the methods using the person mean gave better results.

For unidimensional datasets (Mix1:1), $\overline{\Pi\gamma}$ did not change for any of the methods as the correlation increased. The accompanying standard error

Table 6

Mean of $\Pi\gamma$($\overline{\Pi\gamma}$), and Mean of D^2($\overline{D^2}$) Across 50 Replications, for Different Mixing Configurations, Different Correlations between Traits and Different Percentages of Nonresponse, Fixed at Sample Size 100 and REF-LOWHIGH

(Entries in columns labeled $\overline{D^2}$ are result of multiplication by 1000)

		Correlation Between Traits											
		0				0.24				0.5			
		Percentage of Nonresponse											
		5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
1:0	TW	2.18	16	19.17	140	1.58	13	6.49	129	1.2	11	2.35	92
	CIMS	2.21	16	17.34	141	1.6	14	5.94	131	1.22	12	2.12	96
	PM	2.93	19	>>	168	2.11	17	15.19	149	1.6	14	5.57	108
	IM	10.33	29	>>	278	11.03	29	>>	287	11.18	30	>>	283
	CM	8.71	28	>>	248	8.53	28	>>	252	8.69	28	>>	241
	OM	14.03	37	>>	355	14.85	37	>>	360	15.2	38	>>	359
	TW-E	4.67	26	>>	218	3.34	22	>>	193	2.32	18	>>	133
	CIMS-E	5.07	28	>>	236	3.31	21	>>	198	2.54	18	>>	145
	PM-E	8.01	40	>>	310	4.98	30	>>	258	3.47	24	>>	186

Table 6 (*cont.*)

			Correlation Between Traits											
			0				0.24				0.5			
			Percentage of Nonresponse											
			5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
3:1	TW	1.5	7	12.27	61	1.07	5	4.45	48	0.76	4	0.81	28
	CIMS	2.07	9	6.94	72	1.61	6	3.3	57	1.17	4	0.76	34
	PM	10.68	15	>>	127	6.43	10	>>	98	4.24	8	>>	57
	IM	>>	32	>>	362	>>	30	>>	370	>>	33	>>	387
	CM	>>	43	>>	446	>>	38	>>	469	>>	42	>>	483
	OM	>>	71	>>	756	>>	61	>>	753	>>	63	>>	837
	TW-E	4.63	9	>>	88	3.3	6	>>	67	2.08	5	>>	37
	CIMS-E	8.8	13	>>	98	7.37	9	>>	74	4.15	7	>>	46
	PM-E	>>	28	>>	227	>>	19	>>	181	>>	16	>>	124
1:1	TW	0.64	2	0.34	8	0.66	1	0.26	8	0.63	1	0.23	7
	CIMS	0.67	2	0.2	15	0.71	2	0.18	13	0.69	1	0.15	11
	PM	1.66	3	6.37	21	1.68	2	4.45	15	1.49	1	3.91	11
	IM	>>	42	>>	559	>>	43	>>	556	>>	42	>>	571
	CM	>>	43	>>	495	>>	47	>>	530	>>	45	>>	528
	OM	>>	85	>>	1002	>>	83	>>	942	>>	75	>>	923
	TW-E	1.2	2	5.3	14	1.26	2	5.18	12	1.23	1	4.53	9
	CIMS-E	1.45	3	3.77	19	1.57	3	4.13	17	1.47	2	3.97	16
	PM-E	4.41	12	>>	111	4.69	10	>>	91	4.39	8	>>	69

Table 7

Mean of $\Pi\gamma$($\overline{\Pi\gamma}$), and Mean of D^2($\overline{D^2}$) Across 50 Replications, for Different Mixing Configurations, Different Correlations between Traits and Different Percentages of Nonresponse, Fixed at Sample Size 500 and REF-HIGH

(Entries in columns labeled $\overline{D^2}$ are result of multiplication by 1000)

		Correlation Between Traits											
		0				0.24				0.5			
		Percentage of Nonresponse											
		5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
1:0	TW	1.57	6	3.82	101	1.18	5	1.38	83	0.92	4	0.5	65
	CIMS	1.56	8	3.25	131	1.19	7	1.2	107	0.92	6	0.42	86
	PM	1.89	7	6.84	113	1.42	6	2.55	91	1.11	5	0.99	70
	IM	6.83	11	>>	150	6.65	11	>>	157	6.76	11	>>	155
	CM	4.61	9	>>	132	4.45	8	>>	132	4.54	9	>>	135
	OM	8.02	15	>>	204	7.72	14	>>	208	7.86	15	>>	209
	TW-E	3.52	10	>>	147	2.54	8	>>	109	1.84	6	7.57	77
	CIMS-E	3.61	12	>>	169	2.62	10	>>	128	1.87	7	6.61	93
	PM-E	5.02	15	>>	198	3.52	11	>>	152	2.58	9	>>	110

Table 7 (*cont.*)

		Correlation Between Traits											
		0				0.24				0.5			
		Percentage of Nonresponse											
		5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
3:1	TW	1.03	4	0.65	55	0.8	3	0.23	39	0.62	2	0.09	26
	CIMS	1.53	6	1.31	83	1.2	4	0.62	56	0.99	3	0.3	37
	PM	3.88	8	>>	96	2.84	6	7.97	65	2.08	4	3.21	43
	IM	>>	11	>>	146	>>	11	>>	149	>>	11	>>	156
	CM	>>	16	>>	242	>>	16	>>	226	>>	16	>>	231
	OM	>>	30	>>	404	>>	29	>>	375	>>	27	>>	367
	TW-E	3.36	4	>>	61	2.43	4	17.51	41	1.64	2	5.02	26
	CIMS-E	6.21	7	>>	91	4.76	6	>>	61	3.57	3	>>	39
	PM-E	>>	15	>>	166	19.55	11	>>	112	12.69	8	>>	82
1:1	TW	0.51	1	0.05	16	0.5	1	0.05	13	0.49	1	<<	12
	CIMS	0.53	2	<<	26	0.53	1	<<	21	0.52	1	<<	17
	PM	0.92	1	0.49	9	0.91	1	0.42	7	0.87	0	0.36	5
	IM	11.35	14	>>	217	14.57	14	>>	219	18.89	14	>>	222
	CM	6.38	12	>>	209	8.59	13	>>	216	11.19	13	>>	233
	OM	18.6	29	>>	456	>>	29	>>	430	>>	26	>>	414
	TW-E	1	0	1.01	1	0.99	0	0.95	1	0.98	0	0.94	1
	CIMS-E	1.1	1	1.04	9	1.13	1	1.15	8	1.15	1	1.25	6
	PM-E	2.83	5	>>	61	2.93	4	>>	49	2.86	3	>>	39

Table 8

Mean of $\Pi\gamma$($\overline{\Pi\gamma}$), and Mean of D^2($\overline{D^2}$) Across 50 Replications, for Different Mixing Configurations, Different Correlations between Traits and Different Percentages of Nonresponse, Fixed at Sample Size 500 and REF-MCAR (Entries in columns labeled $\overline{D^2}$ are result of multiplication by 1000)

		Correlation Between Traits											
		0				0.24				0.5			
		Percentage of Nonresponse											
		5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
1:0	TW	1.28	6	1.79	96	1.00	5	0.71	82	0.78	4	0.28	62
	CIMS	1.3	6	1.94	100	1.03	6	0.78	85	0.81	5	0.32	64
	PM	1.47	6	2.93	99	1.16	5	1.17	83	0.91	5	0.49	61
	IM	3.94	6	>>	78	4.04	6	>>	78	3.93	6	>>	77
	CM	3.25	5	>>	62	3.26	5	>>	58	3.2	5	>>	56
	OM	4.41	7	>>	92	4.50	7	>>	91	4.43	7	>>	90
	TW-E	2.94	10	>>	128	2.19	7	13.19	99	1.57	6	4.78	69
	CIMS-E	3.03	10	>>	134	2.26	8	14.93	105	1.68	6	5.59	72
	PM-E	3.96	12	>>	152	2.97	10	>>	122	2.15	7	12.25	80

Table 8 (*cont.*)

			Correlation Between Traits											
			0				0.24				0.5			
			Percentage of Nonresponse											
		5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
3:1	TW	1.01	4	0.56	55	0.76	3	0.22	39	0.62	2	0.1	26
	CIMS	1.35	5	1.6	58	1.08	3	0.79	41	0.91	2	0.39	26
	PM	3.29	6	19.98	63	2.45	4	8.43	43	1.96	3	3.86	29
	IM	>>	7	>>	100	>>	7	>>	102	>>	7	>>	104
	CM	>>	8	>>	87	>>	8	>>	87	>>	8	>>	88
	OM	>>	11	>>	154	>>	11	>>	150	>>	11	>>	148
	TW-E	3.37	5	>>	62	2.27	3	17.83	42	1.7	2	6.02	26
	CIMS-E	5.77	6	>>	71	4.43	4	>>	50	3.49	3	>>	32
	PM-E	>>	10	>>	110	>>	8	>>	78	16.06	6	>>	54
1:1	TW	0.47	1	<<	18	0.47	1	<<	15	0.46	1	<<	13
	CIMS	0.52	1	0.06	17	0.52	1	0.05	14	0.52	1	0.05	11
	PM	0.84	1	0.31	6	0.83	0	0.29	5	0.81	0	0.27	4
	IM	7.12	10	>>	165	10.37	11	>>	165	13.64	11	>>	176
	CM	5.4	8	>>	98	7.78	9	>>	112	10.72	10	>>	131
	OM	10.97	16	>>	239	15.86	16	>>	236	>>	16	>>	241
	TW-E	0.91	0	0.7	1	0.91	0	0.73	1	0.92	0	0.74	1
	CIMS-E	1.08	1	1.31	7	1.12	1	1.37	5	1.17	0	1.62	5
	PM-E	2.65	3	>>	30	2.69	3	>>	25	2.74	2	>>	21

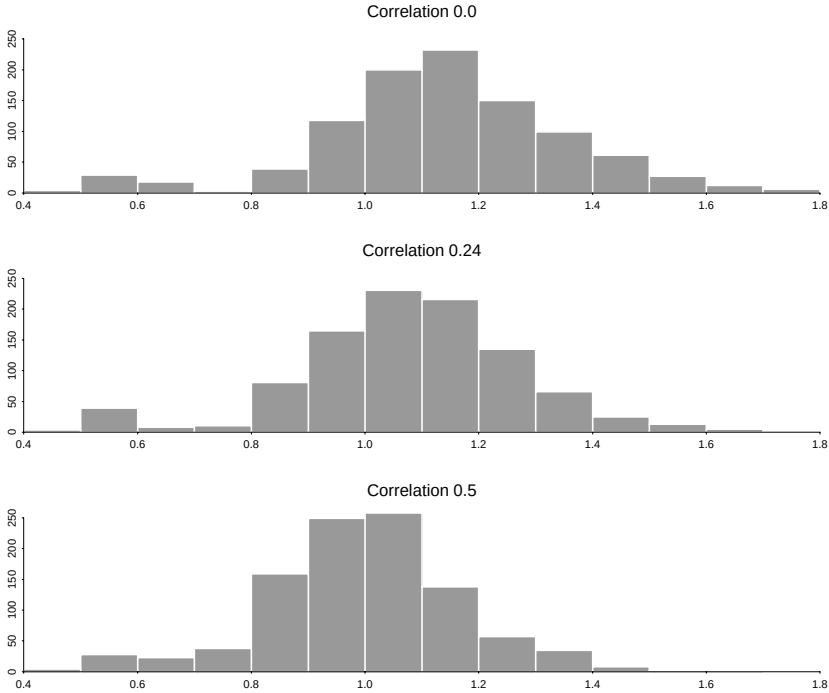


Figure 1

Histograms of 20 Eigenvalues across 50 Replications of Imputation Method TW, for Correlation 0 (Upper Panel); Correlation 0.24 (Middle Panel); and Correlation 0.5 (Lower Panel); and keeping Other Design Factors Fixed at Mix 3:1, Sample Size 100, REF-LOWHIGH, and 20 Percent Missings.

remained approximately equal. $\overline{D^2}$ decreased for TW, CIMS, PM, TW-E, CIMS-E, and PM-E as the correlation between the traits increased.

Nonresponse Mechanisms

For two-dimensional datasets (Mix 1:0 and Mix 3:1), $\overline{\Pi\gamma}$ was closest to 1 for REF-MCAR. $\overline{\Pi\gamma}$ was largest for REF-LOWHIGH. Thus, correspondence between loadings matrices based on complete data and loadings matrices based on imputed data was closest for REF-MCAR and most discrepant for REF-LOWHIGH. This was true for each imputation method.

For all nonresponse mechanisms, of all imputation methods, TW had the lowest $\overline{D^2}$, and TW-E had the second lowest $\overline{D^2}$. Also, for both TW and TW-E, $\overline{D^2}$ did not vary substantially across nonresponse mechanisms. The other ten imputation methods had their lowest $\overline{D^2}$ for REF-MCAR. For all nonresponse mechanisms method OM had the highest $\overline{D^2}$.

For unidimensional datasets (Mix 1:1), $\overline{\Pi\gamma}$ of REF-MCAR and REF-HIGH were both close to 1 for all imputation methods. For REF-LOWHIGH, $\overline{\Pi\gamma}$ was largest.

TW-E had the lowest $\overline{D^2}$ and the lowest $s(D^2)$ under all nonresponse mechanisms. OM had the highest $\overline{D^2}$ and the highest $s(D^2)$.

The results for the nonresponse mechanisms across the varying design cells show that the relative performance of imputation methods did not differ dramatically for different nonresponse mechanisms. Hence, the decision which imputation method to use in practice does not seem to heavily depend on the nonresponse mechanism. These are indications that researchers need not identify the nonresponse mechanism in detail but can use simple indicators, such as the underlying latent trait structure (mixing configuration), for deciding which imputation method to use. For example, if the data are unidimensional method TW-E is a good choice whereas for multidimensional data method TW may be a better option.

Latent Trait Configuration

In general, for all imputation methods, compared with the two-dimensional cases Mix 1:0 and Mix 3:1, the unidimensional case Mix 1:1 had $\overline{\Pi\gamma}$ closest to 1 and the lowest $\overline{D^2}$. In particular, under Mix 3:1 imputation methods had $\overline{\Pi\gamma}$ furthest bounded away from 1 and the highest $\overline{D^2}$. All imputation methods performed best when the data were unidimensional.

Sample Size

Under each of the three nonresponse mechanisms, for all imputation methods $\overline{\Pi\gamma}$ varied little across both sample sizes. However, the standard error of $\Pi\gamma$ at least halved in all cells when sample size increased from 100 to 500. $\overline{D^2}$ and $s(D^2)$ decreased for all imputation methods as the sample size increased. For larger sample size there was less capitalization on chance; this gave better results for sample size 500 than for sample size 100.

EM Algorithms

Tables 9 and 10 show results for EM-loadings, EM-covariances and imputation methods using the person mean. In general, the results for $\overline{\Pi\gamma}$ and $\overline{D^2}$ showed that EM-covariances and EM-loadings better recover the complete data factor loadings matrix than the imputation methods.

Table 9

Mean of $\Pi\gamma$($\overline{\Pi\gamma}$), and Mean of D^2($\overline{D^2}$) Across 50 Replications, for Different Mixing Configurations, Different Correlations between Traits and Different Percentages of Nonresponse, Fixed at Sample Size 100 and REF-LOWHIGH

(Entries in columns labeled $\overline{D^2}$ are result of multiplication by 1000)

		Correlation Between Traits											
		0				0.24				0.5			
		Percentage of Nonresponse											
		5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
1:0	TW	2.35	16	>>	143	1.66	13	6.83	119	1.18	11	1.99	84
	CIMS	2.33	16	18.73	149	1.69	13	5.91	124	1.20	11	1.77	87
	PM	3.20	19	>>	166	2.18	15	16.49	140	1.57	13	4.69	97
	EM cova	1.33	5	4.75	42	1.34	5	4.51	39	1.38	6	4.89	44
	EM load	0.70	4	0.20	31	0.70	5	0.21	31	0.72	6	0.22	33
	TW-E	4.98	26	>>	225	3.38	20	>>	183	2.31	17	>>	125
	CIMS-E	5.20	27	>>	241	3.32	21	>>	186	2.40	17	>>	133
	PM-E	7.66	36	>>	312	5.37	31	>>	235	3.50	25	>>	183

Table 9 (*cont.*)

		Correlation Between Traits											
		0				0.24				0.5			
		Percentage of Nonresponse											
		5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
3:1	TW	1.60	7	16.24	70	1.16	5	3.35	45	0.75	4	1.08	30
	CIMS	2.15	9	8.69	80	1.62	7	2.70	52	1.16	5	1.06	34
	PM	11.42	15	>>	142	7.06	11	>>	91	4.04	8	>>	62
	EM cova	1.12	2	2.26	14	1.14	1	2.37	9	1.13	1	>>	11
	EM load	0.45	2	<<	13	0.45	2	<<	11	0.44	2	<<	12
	TW-E	5.54	10	>>	100	3.45	8	>>	62	2.10	5	>>	41
	CIMS-E	8.06	13	>>	109	6.18	10	>>	70	4.10	7	>>	48
	PM-E	>>	29	>>	286	>>	20	>>	166	>>	16	>>	115
1:1	TW	0.65	1	0.38	9	0.61	1	0.27	8	0.60	1	0.25	6
	CIMS	0.69	2	0.23	15	0.66	2	0.18	12	0.66	1	0.17	11
	PM	1.70	3	7.54	25	1.52	2	4.86	16	1.44	1	4.23	12
	EM cova	1.30	1	4.86	15	1.24	1	3.85	11	1.28	1	4.12	9
	EM load	0.62	1	0.15	13	0.57	2	0.10	13	0.56	1	0.10	10
	TW-E	1.25	2	6.46	17	1.19	2	5.34	12	1.21	1	5.43	10
	CIMS-E	1.43	3	4.26	20	1.45	3	4.04	17	1.45	2	4.53	16
	PM-E	4.51	11	>>	124	4.47	10	>>	88	4.06	7	>>	73

Table 10

Mean of $\Pi\gamma$($\overline{\Pi\gamma}$), and Mean of D^2($\overline{D^2}$) Across 50 Replications, for Different Mixing Configurations, Different Correlations between Traits and Different Percentages of Nonresponse, Fixed at Sample Size 100 and REF-HIGH

(Entries in columns labeled $\overline{D^2}$ are result of multiplication by 1000)

		Correlation Between Traits											
		0				0.24				0.5			
		Percentage of Nonresponse											
		5		20		5		20		5		20	
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
1:0	TW	1.78	12	5.08	128	1.21	11	1.61	109	0.89	9	0.49	91
	CIMS	1.81	14	4.48	162	1.23	13	1.39	134	0.91	12	0.43	112
	PM	2.11	14	8.74	141	1.47	13	3.02	119	1.08	11	0.95	99
	EM cova	1.16	4	1.70	32	1.09	5	1.69	30	1.07	5	1.81	35
	EM load	0.59	4	0.08	35	0.57	5	0.07	34	0.56	5	0.08	38
	TW-E	3.93	20	>>	194	2.51	17	>>	147	1.84	15	7.55	114
	CIMS-E	4.18	23	>>	217	2.72	20	>>	169	1.93	17	7.30	134
	PM-E	5.65	30	>>	254	3.73	27	>>	202	2.58	23	>>	168

Table 10 (*cont.*)

			Correlation Between Traits											
			0				0.24				0.5			
			Percentage of Nonresponse											
		5		20		5		20		5		20		
Mix	Method	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$	$\overline{\Pi\gamma}$	$\overline{D^2}$
3:1	TW	1.10	7	0.74	67	0.80	5	0.32	49	0.64	4	0.12	35
	CIMS	1.58	10	1.83	94	1.19	7	0.86	69	1.06	6	0.46	48
	PM	4.31	12	>>	110	3.07	9	11.20	84	2.29	8	4.51	60
	EM cova	1.07	2	1.39	11	1.04	1	1.49	9	1.05	2	1.34	9
	EM load	0.43	2	<<	13	0.41	2	<<	11	0.41	2	<<	12
	TW-E	3.79	10	>>	78	2.38	7	>>	57	1.76	5	7.49	41
	CIMS-E	6.70	14	>>	112	4.51	10	>>	81	3.94	8	>>	56
	PM-E	>>	26	>>	194	18.74	19	>>	150	14.92	16	>>	117
1:1	TW	0.54	2	0.05	20	0.49	2	0.05	15	0.50	1	<<	14
	CIMS	0.55	3	0.05	31	0.52	2	0.05	23	0.56	2	0.05	19
	PM	1.00	1	0.52	12	0.90	1	0.46	10	0.89	1	0.35	8
	EM cova	1.13	1	1.59	7	1.05	1	1.55	4	1.12	0	2.09	4
	EM load	0.54	2	0.05	21	0.49	2	<<	17	0.51	1	<<	15
	TW-E	1.06	1	1.05	6	1.01	1	1.04	5	1.02	1	0.93	5
	CIMS-E	1.13	2	1.18	17	1.15	2	1.48	13	1.28	2	1.43	10
	PM-E	3.18	10	>>	76	2.88	6	>>	63	2.95	6	>>	47

Percentage of Missing Item Scores

For 5 and 20 percent of missing item scores the order of the missing data methods according to $\overline{\Pi\gamma}$ and $\overline{D^2}$ was the same, with both EM methods performing best.

Correlation Between Traits

For the three correlations between latent traits the EM methods showed the same performance. It may be noted that the performance of the imputation methods varied across correlations between traits.

Latent Trait Configuration

For each of the three latent trait configurations, EM-loadings and EM-covariances performed similarly according to $\overline{\Pi\gamma}$ and $\overline{D^2}$. For both EM methods, $\overline{D^2}$ was lower for Mix 3:1 than for Mix 1:0. $\overline{D^2}$ was lowest for the unidimensional case Mix 1:1.

Nonresponse Mechanism

In general, for nonresponse mechanism REFHIGH EM-covariances had $\overline{\Pi\gamma}$ closer to 1 than EM-loadings. Also, EM-covariances had the lowest $\overline{D^2}$. The results were reversed for nonresponse mechanism REF-LOWHIGH. Here, EM-loadings had $\overline{\Pi\gamma}$ closer to 1 and lower $\overline{D^2}$ than EM-covariances. It may be noted that all differences between methods were modest.

EM-Loadings Versus EM-Covariances

In approximately half of all cells, EM-loadings had a higher $\overline{D^2}$ than EM-covariances. This difference was small for two-dimensional cases and greater for the unidimensional case (Mix 1:1) and nonresponse mechanism REF-HIGH. Probably this difference in $\overline{D^2}$ may be attributed to technical differences between the convergence criteria used.

Discussion

For imputation methods, two main results were found. First, methods using the person mean (that is, TW, CIMS, and PM) in general yielded factor

loadings closer to the complete data factor loadings than methods not using the person mean (that is, IM, CM, and OM). Second, for unidimensional data, methods using the person mean plus a residual error (that is, TW-E and CIMS-E) often performed better than corresponding methods without error. PM-E, however, in general performed worse than PM.

Of the methods not using the person mean, IM gave the best results and will therefore be compared with PM in an effort to better understand the success of methods using the person mean. Unlike IM, PM (and other methods using the person mean) imputes scores based on the observed scores on the *other* items. Since all items are correlated due to a common latent trait structure, PM estimates an imputed score (without error) using information on the latent traits. In contrast, IM imputes a score by taking the mean of the observed scores of the other respondents on the *same* item. Thus, no information about the latent trait structure other than through this particular item is used. This means that for a particular respondent with a missing value on an item, PM uses predictive information about the respondent's latent trait position, whereas IM only uses the group mean on one particular item and thus lacks the predictive power of PM. Consequently, PM and other methods using the person mean yield better estimates than methods not using the person mean and, as a result, the complete data factor loadings are better recovered.

For the unidimensional case we will next explain why TW-E and CIMS-E, which both use the person mean, often yielded factor loadings that were closer to complete data factor loadings than corresponding methods without error (TW and CIMS). Compared with correlations between items based on the complete data, correlations based on data with scores imputed by TW or CIMS tend to be higher because these imputed scores are based on the other items sharing the latent trait structure with the item on which a missing score was observed. This mutual dependence artificially creates stronger association when item scores are imputed than when item scores are all observed. Thus, correlations are too high. Compared with this situation, the addition of a random error component score to the TW or CIMS score has the opposite effect of lowering the correlations in the direction of the complete data correlations. This means that the factor loadings of the complete data and the factor loadings of the data with score-plus-error imputed by means of TW-E or CIMS-E are more similar than the factor loadings of the complete data and the factor loadings of the data with score-without-error imputed by means of TW or CIMS.

Interestingly, PM-E performed worse than PM. It may be noted that PM uses less information for estimating missing scores than TW and CIMS. Method

TW adds to method PM information on the item mean and the grand mean, and method CIMS adds information on all item means. Thus, TW and CIMS may be seen as extensions of PM and are expected to perform better than PM. The reason why adding error has the effect of weakening the performance of PM but not the performance of TW and CIMS is not clear to us.

The goal of imputation is to substitute the missing score by a plausible value. If the imputation method describes the data well, the residual standard error will be small. This means that the imputation is not dominated by its corresponding error. Figure 2 contains six histograms of standard errors for the imputation methods TW, CIMS, PM, IM, CM, and OM, keeping design factors fixed at Mix 3:1, sample size 100, 10 percent missingness, REF-MCAR, and correlation 0 between latent traits. From the histograms it follows that standard errors from imputation methods using the person mean (TW, CIMS, and PM) are the

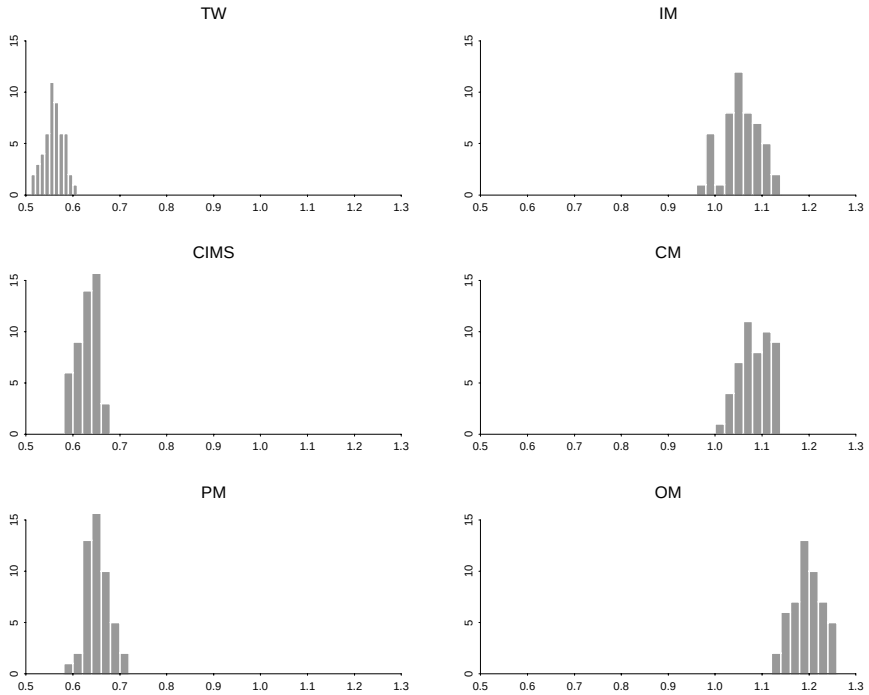


Figure 2

Histograms of Residual Standard Errors for Six Imputation Methods Across 50 Replications of Mix 3:1, Sample Size 100, 10 Percent Missingness, REF-MCAR, and Correlation 0 Between Latent Traits; Residual Standard Error (x - axis) Versus Count (y - axis).

smallest among the imputation methods considered. If an error is added to the imputed score, the methods with the smallest error will give the best results with respect to the measures of discrepancey used here.

For EM methods it can be concluded that both recovered the complete data factor loadings approximately equally well. From a statistical point of view we thus recommend the use of either EM-loadings or EM-covariances for estimating a factor loadings matrix when the rating scale data contain missing scores. Researchers who feel confident to use these relatively complex methods thus are advised to use one of them. From a practical point of view, however, for unidimensional or near unidimensional data, we advise clients who have Likert scale data suffering from item score missingness to use an imputation method depending on the person mean, in particular method TW or method CIMS, and to add a draw from a normal distribution with mean zero and residual variance. For multidimensional data, person mean methods without a draw from the residual error, in particular method TW, give better results with respect to the measures of discrepancy used here because the mean is taken across more than one trait. If these traits are highly correlated, the researcher can proceed as if the data were unidimensional and thus impute scores using method TW-E.

References

Acock, A. C. (1997). Working with missing values. *Family Science Review, 10*, 76-102.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. New York: Wiley.

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988). *The new S language: A programming environment for data analysis and graphics*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Bernaards, C. A. & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research, 34*, 277-313.

Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika, 48*, 269-291.

Cattell, R. B. (1978). *The scientific use of factor analysis in the behavioral and life sciences.* New York: Plenum Press.

Dolan, C. V. (1994). Factor Analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309-326.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika, 44*, 409-420.

Huisman, M. (1998). *Item nonresponse: Occurrence, causes and imputation of missing answers to test items*. Leiden: DSWO Press.

Kelderman, H. & Rijkes, C. P. M. (1994). Loglinear Multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149-176.

Lee, Sik-Yum (1986). Estimation for structural equation models with missing data. *Psychometrika, 51*, 93-99.

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 149-158.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Little, R. J. A. & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research, 18*, 292-326.

Little, R. J. A. & Schenker, N. (1995). Missing data. In G. Arminger, C.C. Clogg, & M.E. Sobel (Eds.), *Handbook of statistical modeling in the social and behavioral sciences* (pp. 39-75). New York: Plenum Press.

Liu, C. & Rubin, D. B. (1998). Maximum likelihood of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica, 8*, 729-747.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer-Verlag.

Muraki, E. & Carlson, J. E. (1995). Full-information factor analysis of polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.

Muthén, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.

Muthén, B. & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: a note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431-462.

Niesing, J. (1997). *Simultaneous component and factor analysis methods for two or more groups: a comparative study*. Leiden: DSWO Press.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56*, 241-254.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473-489.

Rubin, D. B. & Thayer, D. T. (1982). EM algorithms for factor analysis. *Psychometrika, 47*, 69-76.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

SPSS (1989). *SPSS* [Computer software]. Chicago: Author.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Tanner, M. A. (1996). *Tools for statistical inference. Methods for the exploration of posterior distributions and likelihood functions*. New York: Springer-Verlag.

Ten Berge, J. M. F. (1977). *Optimizing factorial invariance*. Doctoral thesis, University of Groningen.

Accepted October, 1999.

Appendix

The EM algorithm for estimating a covariance matrix suffering from non-response under multivariate normality is described in detail in Little and Rubin (1987) and Schafer (1997). We briefly outline the procedure followed by Splus routines to carry out the actual calculations. The EM algorithm relies on the following results from multivariate analysis. For a proof we refer to Anderson (1984, p. 37).

Theorem

Let the components of $\mathbf{X}$ be divided into two groups composing the subvectors $\mathbf{X}_1$ and $\mathbf{X}_2$. Suppose the mean $\boldsymbol{\mu}$ is similarly divided into $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and suppose the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{X}$ is divided into $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, and $\boldsymbol{\Sigma}_{22}$, the covariance matrices of $\mathbf{X}_1$, of $\mathbf{X}_1$ and $\mathbf{X}_2$, and of $\mathbf{X}_2$ respectively. Then the distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$

Suppose that the item responses for N persons are $(\mathbf{Y}_1, ..., \mathbf{Y}_N)$. For each respondent, subdivide the scores into an observed part and a part with missings, thus $\mathbf{Y}_i = (\mathbf{Y}_{obs,i}; \mathbf{Y}_{mis,i})$. Let $\boldsymbol{\mu}^t$ and $\boldsymbol{\Sigma}^t$ denote the parameter estimates for the mean and the covariance matrix, respectively, at cycle t.

The EM algorithm consists of two steps. In the t-th cycle of the E step, for each person separately, the mean and covariance matrix of $\mathbf{Y}_{mis,i}$ are calculated given $\mathbf{Y}_{obs,i} = \mathbf{y}_{obs,i}$ and the parameter estimates at cycle t. Thus, using the theorem above with $\mathbf{X}_1$ substituted by $\mathbf{Y}_{mis,i}$ and $\mathbf{X}_2$ substituted by $\mathbf{Y}_{obs,i}$, the unobserved values $\mathbf{Y}_{mis,i}$ are replaced by their expectations,

$$E(\mathbf{Y}_{mis,i}|\mathbf{Y}_{obs,i} = \mathbf{y}_{obs,i}, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = \boldsymbol{\mu}^t_{mis} + \boldsymbol{\Sigma}^t_{mis,obs}\boldsymbol{\Sigma}^{-1\ t}_{obs,obs}(\mathbf{Y}_{obs} - \boldsymbol{\mu}^t_{obs}).$$

The covariance matrix of $\mathbf{Y}_{mis,i}$ is calculated as follows,

$$\text{Var}(\mathbf{Y}_{mis,i}|\mathbf{Y}_{obs,i} = \mathbf{y}_{obs,i}, \mathbf{m}^t, \mathbf{S}^t) = \mathbf{S}^t_{mis,mis} - \mathbf{S}^t_{mis,obs}\mathbf{S}^{-1\ t}_{obs,obs}\mathbf{S}^t_{obs,mis}. \quad (5)$$

The (co)variances of $\mathbf{Y}_{obs,i} = \mathbf{y}_{obs,i}$, and the covariances between $\mathbf{Y}_{mis,i}$ and $\mathbf{Y}_{obs,i} = \mathbf{y}_{obs,i}$, are equal to zero. Note that Equation 5 only depends on the pattern of nonresponse for respondent i; thus calculation of the variances for all patterns of nonresponse and then choosing the right one, would suffice. However, for 20 items, there would be

$$\sum_{i=0}^{20}\binom{20}{i}=2^{20}=1048576$$

possible patterns. Therefore, to ease programming, we prefer to calculate the variance for each person separately, even though that could result in calculation of the variance across the same pattern of missingness more than once.

The M step consists of updating the parameter estimates for $\boldsymbol{\mu}^t$ and $\boldsymbol{\Sigma}^t$. First, $\boldsymbol{\mu}^t$ is updated to $\boldsymbol{\mu}^{t+1}$ by calculating the mean across all persons including the updated values for $\mathbf{Y}_{mis}$. Next, $\boldsymbol{\Sigma}^t$ is updated to $\boldsymbol{\Sigma}^{t+1}$ as follows,

$$\sum\nolimits_{jl}^{t+1}=\frac{1}{N}\sum_{i=1}^{N}\left[\left(\mathbf{Y}_{ij}-\mathbf{m}_j^{t+1}\right)\left(\mathbf{Y}_{il}-\mathbf{m}_j^{t+1}\right)+Var\left(\mathbf{Y}_{mis,i}|\mathbf{Y}_{obs,i}=\mathbf{y}_{obs,i},\mathbf{m}^t,\Sigma^t\right)_{jl}\right].$$

The iterations continue until convergence. We based convergence on the maximum relative change of the parameters. Convergence was obtained when the relative change of all parameters did not exceed .0001. We used as starting values the parameter estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ under listwise deletion.

The following Splus programs produce an EM estimated covariance matrix. The main routine is called EMcov and needs a data matrix **A** (not a data frame) and a convergence criterion (standardly set at 0.0001) as input. Two subroutines, condivar and rel.change, are called by EMcov. The output is the EM estimated covariance matrix, $sigmat1, a corresponding mean vector, $mut1, and the number of iterations, $iterations, needed by the algorithm before convergence. The algorithms have been tested thoroughly but the authors accept no liability whatsoever for the actual use of the routines. Disclaimer: the authors accept no responsibility for the correctness or usability of the results of this software.

```
EMcov<-function(A,max.tol = 0.0001){
   m <- is.na(A)
   k <- sum(m)
   dA <- dim(A)
   if(k == 0)return(A)
   nam <- names(A)
   if[is.null(nam)]{
      nam<-as.character(1:length(A))
      names(A)<-nam
   }
```

```
   Bnew <- apply(A,2,impute,mean)
   toler <- 1
   mut1 <- apply [na.omit(A),2,mean]
   sigmat1 <- var[na.omit(A)]
   iterations <- 0
   while(toler > max.tol){
      iterations <- iterations + 1
      B <- Bnew
      out <- condivar(B,m,mut1,sigmat1)
      Bnew <- out$x
      sigmat <- out$sigmat
      sigmaold <- sigmat1
      muold <- mut1
      mut1 <- apply(Bnew,2,mean)
      Bminmu <- sweep(Bnew,2,mut1,"-")
      sigmat1 <- (t(Bminmu)
      apply(sigmat,c(1,2),sum))/dA[1]
      toler<- rel.change(muold,mut1,sigmaold,sigmat1
   }
   return(sigmat1,mut1,iterations)
}
condivar < -function(x,m,mu,sigma){
   xrow <- nrow(x)
   xcol <- ncol(x)
   sigmat <- array(0,c(xcol,xcol,xrow))
   for(iin1:xrow){
      if(sum(m[i,])>0){
            miss<-seq(xcol)[m[i,]]
            obs <- seq(xcol)[!m[i,]]
            sigmaobs <- as.matrixginversesigma
            [obs,obs]))
            sigma12 <- t(as.matrix(sigma[obs,miss]))
            if(length(obs) == 1)
                  sigma12 <- t(sigma12)
            sigmamiss <- sigma[miss,miss]
            x1 <- mu[miss]
            x2 <- x[i,obs]
            mu2 <- mu[obs]
            x[i,miss] <- x1+sigma12
            sigmat[miss,miss,i] <- sigmamiss-
            sigma12
                  sigmaobs
      }
   }
```

```
   return(x,sigmat)
}
rel.change <- function(x1,x2,y1,y2){
   m <- (abs(x2-x1)>1e-05)
   ch1 <- max(abs((x2[m]-x1[m])/x1[m]))
   m <- (abs(y2-y1)>1e-05)
   ch2 <- max(abs((y2[m]-y1[m])*(1/y1[m])))
   change <- max(ch1,ch2)
   if(is.na(change))
   change <- 0
   change
}
```